A Development of Preprocessing Models for Toll Collection System Data

Lee Hyun-seok Namkoong Seong

Transportation Research Division

Korea Expressway & Transportation Research Institute

50-5, Sancheok-ri, Dongtan-myeon, Hwaseong-si, Kyeonggi-do, Korea

lhsykm@ex.co.kr, jake@ex.co.kr

ABSTRACT

TCS Data imply characteristics of traffic conditions. However, there are outliers in TCS data, which can not represent the travel time of the pertinent section. if these outliers are not eliminated, travel time may be distorted owing to these outliers. Various travel time can be distributed under the same section and time because the variation of the travel time is increase as the section length is increase, which make difficult to calculate the representative of travel time. Accordingly, it is important to grasp travel time characteristics in order to compute the representative of travel time using TCS data.

In this study, after analyzing the variation ratio of the travel time according to the link distance and the level of congestion, the outlier elimination model and the smoothing model for TCS data were proposed. The results show that the proposed model can be utilized for estimating a reliable travel time for a long-distance path in which there is a variation of travel times from the same departure time, the intervals are large and the change in the representative travel time is irregular for a short period.

1. Introduction

Research Background and Purpose

Recently, research on travel time on national roads have been actively made by using data on sections such as the DSRC, AVI, and GPS. As for the expressways in Korea, however, data on stops using Vehicle Detector Stations (VDSs) are still used to estimate travel time from the first stop to the last stop. There are a few issues of estimating travel time by using a stop detector. For spatiotemporal statistics on the rate of the detector, harmonic mean is calculated by considering effects on the volume of traffic and distance. However, on sections where the installation gap of detectors is not dense, speed is not estimated properly in a crowded or traffic flow transition state. In addition, the travel time, which will be experienced by a vehicle while moving on the target section, is supposed to be the time when a vehicle passes each detector rather than the passing speed detected by each detector at the current time. Currently, however, the travel time denotes an instantaneous travel time calculated by simultaneously collecting data on the passing speed at a stop at the current time from the VDS on each section of

expressways. The time is just a value showing the road situation at the current time rather than a travel time on a route.

Thus, data on sections should be used to estimate travel time that can show changes in traffic situation. Under current circumstances, data on Toll Collection System (TCS) are the reliable data on sections and can be obtained without additional installation costs. The data can show the traffic situation experienced by drivers on a route when travel time is estimated. The number of the TCS data collected from the 262 operation centers of the Korea Expressway Corporation nationwide is about 3,200,000 on average per day. From the traffic data, information on travel time of vehicles on an expressway between the start operation center and arrival operation center can be obtained. The TCS data are very helpful for the estimation of travel time since they are true values showing dynamic characteristics such as the traffic situation experienced by each driver between the first stop and the last stop during driving. Despite, the data have not been used to estimate travel time, but have been used as sales settlement.

In order to offer reliable traffic information, a pre-treatment process for extracting significant traffic data from the observed data as well as the level of the data is crucial. The reason for this is that any and all traffic data are provided after being processed from observed data or being processed by using algorithms. In this paper, a pre-treatment method is proposed to extract significant traffic data by properly processing and cleansing the TCS source data.

Research Details and Methodology

The TCS source data imply traffic characteristics that can reflect traffic situation on a section to some degree. However, the TCS data contain the travel time, overspeed, frequent road line changes, and stoppage on a road shoulder of vehicles that have stayed at service areas for a long amount of time or that have arrived much later than other vehicles due to the stoppage on a road shoulder caused by a vehicle breakdown even if they started at the same time. The data also contain the travel time of vehicles that have arrived much earlier than other vehicles owing to the use of bus lanes. The data do not represent a travel time on the section; thus, if the abnormal values are not eliminated but are clustered, a very different travel time may be estimated due to the abnormal values.

In particular, the distribution of travel time increases on long-distance sections. Even on the same section or at the same time, a wide range of travel time is distributed. As a section gets longer, the travel time varies. On the section between Seoul and Daejeon, it is difficult to calculate an appropriate representative value at each start time. If a variation in travel time is high, it will be difficult to examine a travel pattern on a section from the TCS data. In order to calculate a representative value of travel time based on the TCS data, the variation characteristics of travel time should be found out.

A variation in travel time should be applied differently depending on the length of a section. For instance, if a variation in travel time is 10 minutes, it will be able to represent a change in

traffic situation on the section between Seoul and Giheung; however, it will not be able to do so on the section between Seoul and Daejeon. In addition, in order to reflect a sudden variation in travel time at a time when a transition occurs from congestion to non-congestion or vice versa, a variation in travel time should be applied differently even on the same section depending on the congestion level.

A representative value from which an outlier is eliminated at each aggregate time may show an abnormal value in terms of the daily pattern of travel time. That is, travel times on some section on a day correlate under the congestion circumstances on the section, and show a constant pattern due to the correlation. Thus, the representative values, which are out of the pattern and corresponds to noises, should go through a smoothing process in order not to distort the significant pattern owned by the TCS data.

In this paper, the pre-treatment method for the TCS data is improved, and a significant travel time whose spatiotemporal travel pattern can be found from the TCS source data is extracted by considering the length of a section and variations in travel time depending on traffic time.

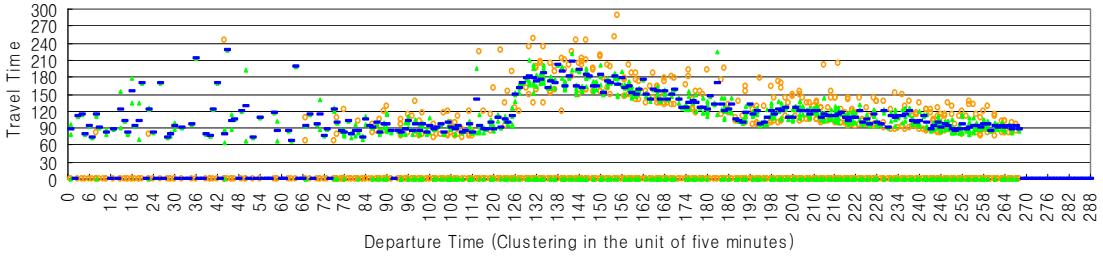
2. Review of Existing Literatures

Most of the traditional pre-treatment algorithms use arithmetic mean to calculate the representative value of travel time. The standard deviation of observed data is calculated, and the upper and lower effective ranges of data are limited. If the variation width of data is large, a few very abnormal values may increase the standard deviation and distort the mean value. In addition, statistical methods are used in most of the researches; however, in statistical methods such as the boxplot method and Shapiro-Wilk test, while deviative data or out-of-center data are treated as an outlier, the normality of distribution is assumed. Traffic data such as the time taken to drive on a section have right-deviative distribution rather than normal distribution. There are many issues in the application of the traditional statistical methods.

Statistical Method

An outlier is eliminated in a statistical method in the sequence below: The travel time of a vehicle on a section is estimated. Then values exceeding either the upper-limit value or lower-limit value are eliminated. In this case, the upper-limit value used by Gang Jin-gee et al. (2002) is the travel time more than twice as fast as the design speed of the section. The lower-limit value is a travel time when a vehicle is driven on the section at the speed of 10km/h. If the number of travel time values reaching the lower-limit value exceeds 50% of the travel time values on the section, values exceeding 68% of confidence interval will be deemed to be outliers and be eliminated. Figure 2.12 shows the effective data, which remain after being eliminated in an outlier elimination method, and the data average. As shown in Figure 1, this method has a weakness as follows: effective data with a relatively large width remain and a

mean value is high overall. Especially, a mean value is greatly distorted by a small number of abnormal values when the number of observed vehicle is low or traffic is congested.



<Figure 1> Elimination of Abnormal Values in a Statistical Method

TransGuide Algorithm

TransGuide is a traffic control system for expressways, which is used in San Antonio, USA. The link travel time between the continuous AVI readers is estimated by the moving average algorithm, which automatically eliminates travel time values exceeding the range set by the user within the collection cycle.

$$tt_{ABi} = \frac{\sum_{i=1}^{|Stt_{ABi}|}(t_{Bi} - t_{Ai})}{|Stt_{ABi}|}$$

Stt_{ABi} = Number of Effective Vehicles that Have Driven Section AB for t

t_{Ai} = Time when Individual Vehicle i Passes Location A

t_{Bi} = Time when Individual Vehicle i Passes Location B

t = Collection Cycle of Travel Time

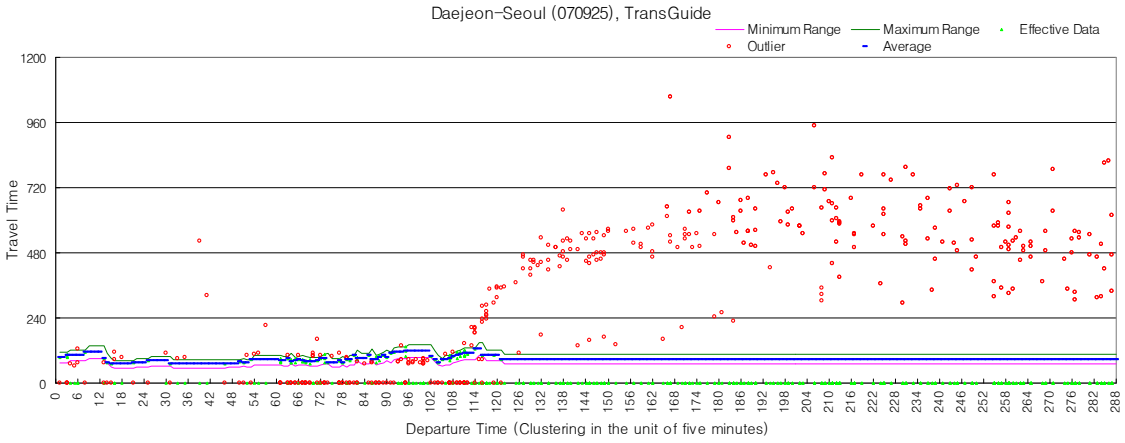
t_w = Moving Average Window

l_{th} = Travel Time Parameter (0.2)

tt_{ABi} = Mean Travel Time of Effective Vehicles at Time t

tt'_{ABi} = Mean Travel Time of Effective Vehicles at Time t-1

In the TransGuide algorithm, the main parameters are t_w and l_{th}. t_w denotes the collection cycle that should be considered to estimate the mean travel time of the current cycle. t_w is typically set to two minutes, and only vehicles that have arrived Location B between Current Time t and Time Range (t-t_w) are deemed to be effective vehicles. l_{th} denotes a difference between the effective travel time of the previous collection cycle and that of the current one, and is set to 0. If the value is different from the estimated value of travel time in the previous stage by greater than 20%, it will be deemed to be an abnormal value.



<Figure 2> Elimination of Abnormal Values by TransGuide Algorithm

In the TransGuide algorithm, in order to determine whether the current collection cycle is normal or not, only the mean travel time of the previous collection cycle is used to set an effective range to the minimum and maximum values, and the mean value of the effective data included in the effective range is calculated. In this algorithm, the number and values of the effective travel time collected from the previous cycle affect the data of next cycle.

If the initially collected data have a value of deviated size, the true value will be likely to be eliminated and the outlier will remain. If the number of clustered data is small, the representative value will be highly likely to be distorted under the great influence of the outlier. In addition, if a travel time sharply increases in special transportation periods and is different from the previous time period by greater than 20%, the travel time will not be deemed to be within the effective range but be deemed to be an outlier. That is, since l_{th} is set to a constant value, a variation in travel time depending on the congestion or section length cannot be considered.

Transmit Algorithm

Transmit is a traffic control system used in New York City and New Jersey. Its method for the estimation of travel time using AVI data is basically similar to TransGuide. In order to estimate the current travel time, moving average is not used but a smoothing method for 15 minutes of collection cycle is used. That is, the data on the smoothed previous cycle with the outlier left is used.

$$tt_{ABk} = \frac{\sum_{i=1}^{n_k}(t_{Bi} - t_{Ai})}{n_k}$$

tt_{ABk} = Average Travel Time on Section AB at Time k

t_{Ai} = Detection Time at Location A

t_{Bi} = Detection Time at Location B

n_k = Number of Vehicles Observed at Time k

The travel time of the collection cycle is averaged as described above. In order to calculate the updated mean travel time, smoothing is performed by using the data on the same day of the week and the same time. The smoothed mean travel time of the current collection cycle is calculated by using the value of smoothed travel time during the previous collection cycle.

$$tth''_{ABk} = (\alpha) \times tth_{ABk} + (1+\alpha) tth''_{ABk-1}$$

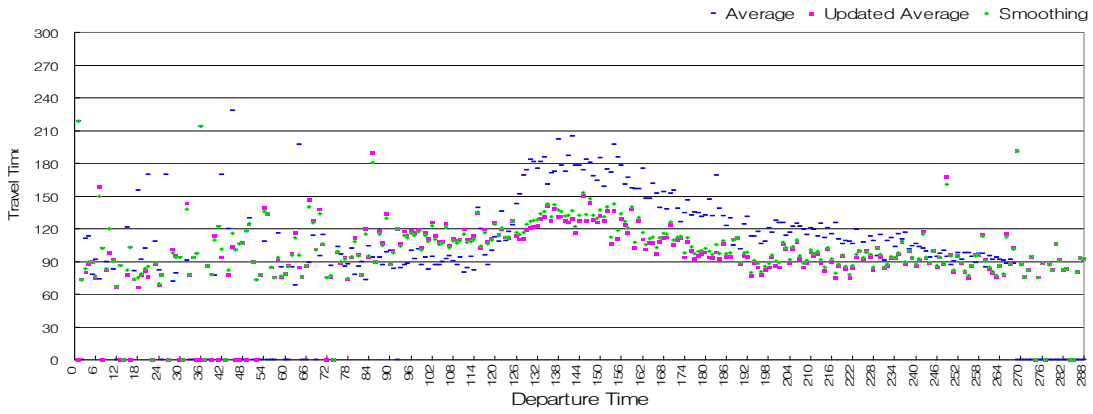
tth_{ABk} = Previous Smoothed Value at Time k

tth''_{ABk} = Smoothed Value at Time k

tth''_{ABk-1} = Smoothed Value at Time $k-1$

α = Smoothing Coefficient (0.1)

The estimation basis of smoothing coefficient in the Transmit algorithm is unclear. If any incidents do not occur, 10% of weight will be given to the updated value of mean travel time. If an incident occurs, 0% of weight will be applied. Thus, in this type of smoothing, data on travel time in the event of an incident are not included in moving average. Only typical repetitive travel times are contained in the history database. As shown in Figure 2.14, there are limitations of estimating travel time by smoothing only in an analysis section where many outliers much different from normal values occur. Thus, it is difficult to apply this algorithm to sections with abnormal and continuous characteristics.



<Figure 3> Elimination of Outliers by the Transmit Algorithm

3. Model of Outlier Elimination

Basic Formula of Outlier Elimination

If a travel time is computed at an interval of five minutes to estimate the representative value, a time-series variation in travel time will be irregular and will contain an outlier with a high variation within the same collection interval. In general, in order to eliminate the outlier within the same collection interval, the distribution of travel time is assumed to be normal distribution. Then, the observed value ($\mu \pm n\sigma$), which is much higher or lower than the mean value by much greater than the standard deviation, is deemed to be the outlier. However, since a standard deviation is calculated by finding the square of a distance from the mean value, an outlier much different from the mean value may rather affect the standard deviation of all the observed values.

The result of applying a model of outlier elimination using median absolute deviation when calculating the representative value of travel time within the collection interval is as follows:

$$MAD = median \mid x_i - x_{med} \mid$$

MAD: Median Absolute Deviation

x_i : i th Travel Time Observed within the Collection Interval

x_{med} : Median Value of Travel Time Observed within the Collection Interval

The approximation of the MAD by the standard deviation of normal distribution is as follows:

$$\hat{\sigma} = K \bullet MAD$$

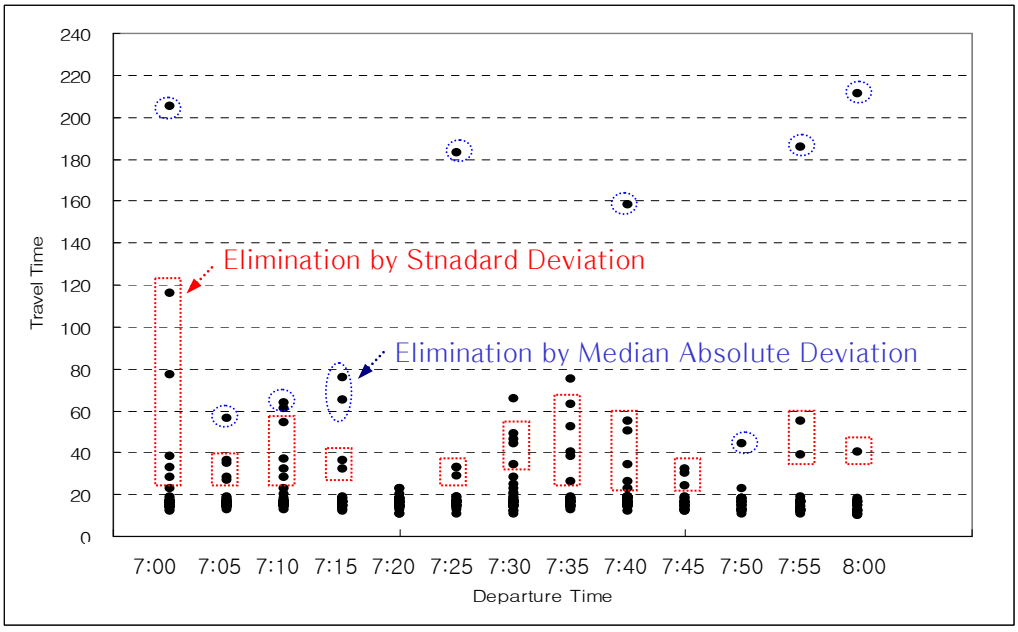
K is the adjustment coefficient of making the MAD the same as the standard deviation of normal distribution. Since the MAD is a distance between the first quartile and the second quartile, the probability formula below is established:

$$\frac{1}{2} = P\left(|X - \mu| \le MAD\right) = P\left(\left|\frac{X - \mu}{\sigma}\right| \le \frac{MAD}{\sigma}\right)$$

$$= P\left(|Z| \le \frac{MAD}{\sigma}\right)$$

In a normal distribution of $\hat{\sigma} = 1$, $\frac{MAD}{\sigma} = \phi^{-1}(3/4) \approx 0.6745$. Thus, the following is established: $\hat{\sigma} = 1.4826 \bullet MAD$, $z_i^{MAD} = \frac{|x_i - x_{med}|}{1.4826 \bullet MAD}$

The MAD of the individual observed values of travel time is approximated by z_i^{MAD}, a probability formula following standard normal distribution. The result is compared to the elimination variable already set. In the event of $z_i^{MAD} > z_{cut}$, the result value is deemed to be an outlier. In this case, since probability is 99% in the event of z=3 in standard normal distribution, the typical setting is z_{cut}=3.



<Figure 4> Comparison of the Methods for Outlier Elimination

For any and all vehicles that left the Seoul operation center at AM 7:00 to AM 8:00 on February 2, 2009 and have arrived Anseong, the total of travel times was computed in the unit of five minutes based on the start time and the outliers were eliminated by using the standard deviation and median absolute deviation. The result is shown in Figure 4. If the standard deviation is used, the observed value with 103 minutes of observed value among the vehicles that started at AM 7:00 will distort the deviation of all the observed values and observed values with 63 and 78 minutes of travel time will not be eliminated as outliers. Thus, the representative value of travel time will be 25.2 minutes. However, if the median absolute deviation is used, all the observed values with travel time of 26 minutes or longer and 16 minutes or shorter will be eliminated as outliers. Thus, the representative value of travel time will be 21.0 minutes. That is, if the median absolute deviation is used, outliers out of herd driving within the same aggregate interval will be able to be eliminated efficiently.

Setting of z_{cut}, Elimination Variable

As a travel distance gets longer, a variation in the travel time of vehicles that start at the same time increases. In order to compare a variation in travel time within the same aggregate interval depending on the travel distance, Coefficient of Variation (CV) in travel time, which is calculated by dividing the deviation of travel time by the representative value of travel time, is used.

$$CV_t = \sigma_t \Big/ \mu_t$$

CV_t : CV of Travel Time of Vehicles that Started at Time t

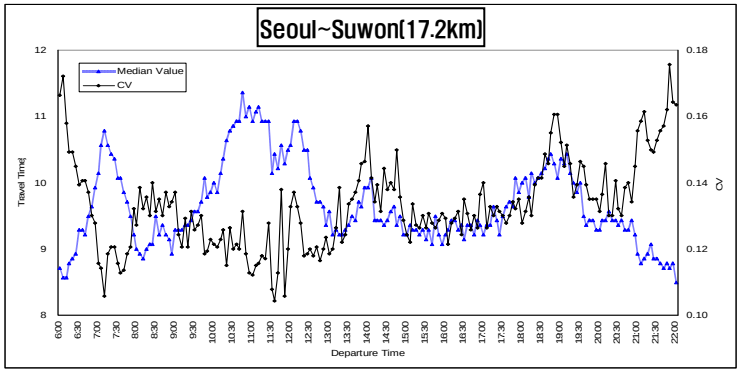
μ_t : Representative Value of Travel Time of Vehicles that Started at Time t

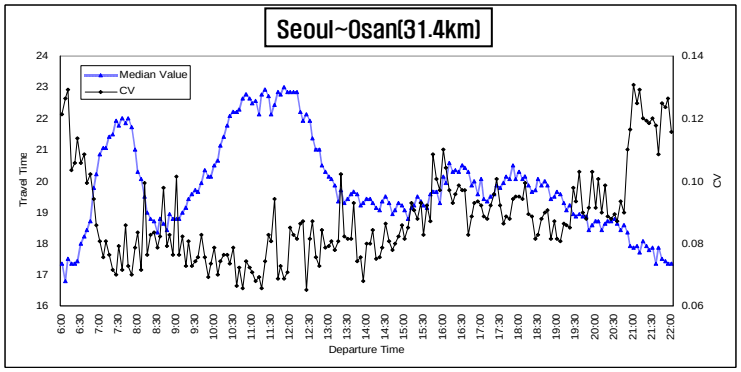
σ_t : Deviation of Travel Time of Vehicles that Started at Time t

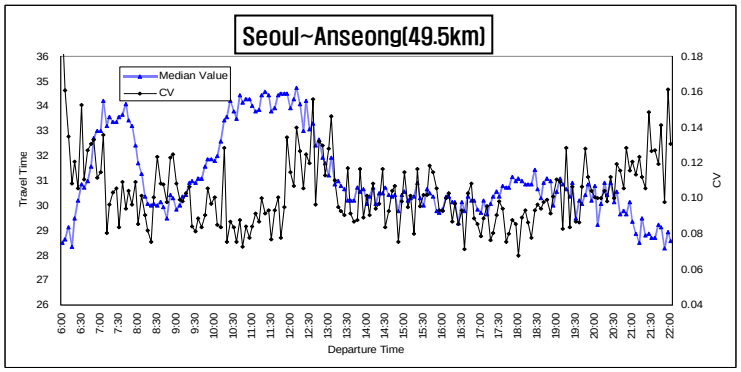
$$= \sqrt{\frac{\sum_{i-1}^{n} \left(T_t^i - \mu_t\right)^2}{n-1}}$$

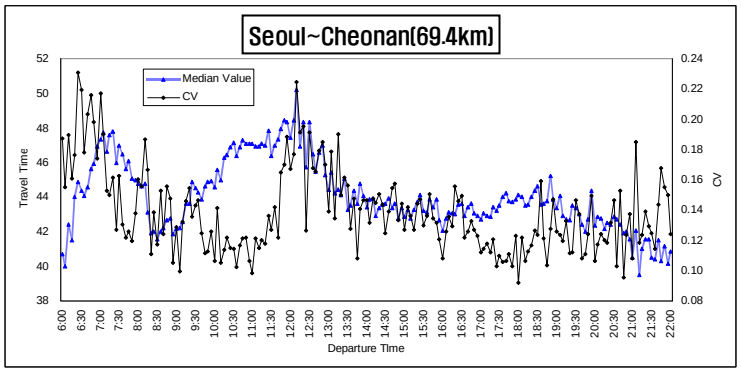
T_t^i : Travel Time of the i th Vehicle that Started at Time t

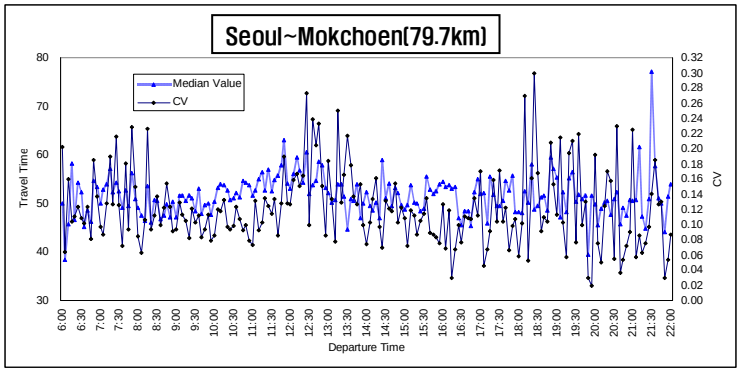
n : Total Number of Vehicles that Started at Time t


Seoul~Suwon(17.2km)
Median Value
CV
Travel Time
CV
Departure Time


Seoul~Osan(31.4km)
Median Value
CV
Travel Time
CV
Departure Time


Seoul~Anseong(49.5km)
Median Value
CV
Travel Time
CV
Departure Time


Seoul~Cheonan(69.4km)
Median Value
CV
Travel Time
CV
Departure Time



<Figure 5> Variations in CVs at Each Hour

In order to analyze the CV of travel time in the event of congestion or non-congestion depending on the section length, the travel times were averaged at each start time in the unit of five minutes and the median and CV of travel time were calculated as shown in Figure 5. From the figure, it is shown that one peak hour at around AM 7:00 and congestion hours of AM 10:00 to noon occur in Seoul. As a travel time increases, the CV of travel time decreases. At congestion hours, the CV ranges from 0.07 to 0.11, while the CV ranges from 0.14 to 0.20 at non-congestion hours.

Especially, on the sections between Seoul and Suwon and between Seoul and Osan, a negative relationship between travel time and the CV is shown. On the long-distance sections between Seoul and Anseong and Seoul and Cheonan, the CV against travel time starts to change much. On the section between Seoul and Mokcheon, the CV is high irregularly regardless of start times or travel times. Considering changes in the CV depending on section lengths, the distance of a single section should not exceed 70km when a travel time is calculated by using the TCS.

When the median absolute deviation is used, z_{cut} =3 was set by default. However, as the CV of travel time increases, a variation in travel time is high. Thus, z_{cut} needs to be adjusted depending on the level of the CV. In this paper, provided that a variation of up to 20% is accepted when a travel time is estimated, The CV is set as follows: z_{cut} =3 by default at σ/μ =0.1. As σ/μ increases, z_{cut} gradually decreases. Thus, z_{cut} depending on the CV is as follows:

$$z_{cut} = \frac{0.1 \times 3}{CV} \text{ at } CV \times z_{cut} = 0.1 \times 0.3$$

CV	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.20
z_{cut}	3.00	2.73	2.50	2.31	2.14	2.00	1.88	1.76	1.67	1.58	1.50

<Table 1> z_{cut} Depending on the CV

4. Smoothing Model

Basic Model of Smoothing

The elimination of outliers using the median absolute deviation means the elimination of individual vehicles whose travel time is abnormal within five minutes of aggregate interval. If the number of vehicles is small within the aggregate interval even if the outlier of each vehicle is eliminated, the representative value at an interval of five minutes may change sharply within a short period of time. In the method of using median absolute deviation, if two groups of vehicles performs herd driving, outliers cannot be identified properly.

A total of four vehicles started on the section between Osan and Cheonan at AM 8:20 on January 23, 2009. The travel time of each vehicle is 20 minutes, 22 minutes, 35 minutes, and 39 minutes. If data on travel times show 20 minutes, 22 minutes, and 35 minutes, z_i^{MAD} of 35 minutes of travel time will be 4.4 and be deemed to be an outlier. In this case, the representative value of travel time is 21 minutes. If data on travel times show 20 minutes, 22 minutes, 35 minutes, and 39 minutes, z_i^{MAD} of 35 minutes and 39 minutes will be 0.6 and 0.9 respectively. In this case, these values are not deemed to be outliers, and the representative value of travel time is 29 minutes. Considering that the representative value of travel time of vehicles that started at AM 8:15 and AM 8:25 is 22 minutes and 24 minutes respectively, it is reasonable to deem 35 minutes and 39 minutes of travel times of vehicles that started at AM 8:20 to be outliers. In the method of removing outliers within the aggregate interval using median absolute deviation, whether the travel times are outliers or not cannot be determined. Thus, smoothing considering variations in travel times before and after the aggregate interval is required after outliers are eliminated within the aggregate interval.

The travel times aggregated in the unit of five minutes based on the start times shown in the TCS data show time-series variations. The variation of travel time at a specific hour shows a difference from the representative value of travel time at the previous hour.
The basic model of smoothing is as follows:

$$\hat{t}_n = \hat{t}_{n-1} + k(t_n + \hat{t}_{n-1})$$

$\hat{t}_n$: Representative Value of Smoothed Travel Time at Hour n

$\hat{t}_{n-1}$: Representative Value of Smoothed Travel Time at Hour n-1

t_n : Representative Value of Travel Time Observed at Hour n

k : Smoothing Constant (0~1), $k = e^{|t_n - \hat{t}_{n-1}|}$

The representative value of smoothed travel time at Hour n is calculated by adding the variation of travel time at Hour n to the smoothed value of the previous hour. If the amount of variation is reasonable, the amount will be accepted and be set to the observed value of travel time as close as possible. If the amount of variation is high, the observed value of travel time will be highly likely to be an outlier and be set to the smoothed value of the previous hour as close as possible. In this case, k, smoothing constant should be applied differently depending on the section length. The reason for this is that the variation distribution of travel time varies depending on section lengths.

By default, k has a value ranging from 0 to 1. When k equals 0, $\hat{t}_n = \hat{t}_{n-1}$ and the representative value of travel time at Hour n is deemed to be an outlier. In addition, the smoothed value of the previous hour is used. In the event of k=1, $\hat{t}_n = t_n$ and the representative value of travel time at Hour n is recognized completely. In the event of 0<k<1, $\hat{t}_n = \hat{t}_{n-1} + k(t_n - \hat{t}_{n-1})$, and the amount of variation at Hour n is applied differently depending on the k value. As the net variation of travel time gets greater, the observed value at the current time is highly likely to be an outlier. Thus, k is set to a low value in order to ensure better smoothing and to apply the amount of variation less.

Setting of k, Smoothing Constant

1) Introduction of r, Distance Coefficient

The variation distribution of travel time varies depending on section lengths. When the start time is different by five minutes, the travel time is different by 20 minutes. The meaning is different between the Seoul-Suwon section (17.2km) and the Seoul-Cheonan section (69.4km). That is, the same amount of variation is more likely to be an outlier on the Seoul-Suwon section than on the Seoul-Cheonan section whose length is longer. The k value showing the level of smoothing should be applied differently depending on section lengths.

In order to differentiate the effects of the same amount of variation depending on section lengths, Distance Coefficient r is introduced to Smoothing Constant k, and the constant is converted into the following:

$$k = e^{\frac{1}{r}|t_n - \hat{t}_{n-1}|}$$

The upper-limit value of r is set to 1 while that lower-limit value of r is set to 3 by introducing a logistic function. Depending on section lengths, the formula below is created so that r can be set to 1 to 3:

$$r = \frac{{}^{r}\max - {}^{r}\min}{1 + \frac{{}^{r}\max - {}^{r}\min}{{}^{r}\min} e^{-m(d-1)}} + 1 = \frac{2}{1 + 2e^{-m(d-1)}} + 1$$

In order to determine the shape of a logistic curve showing a variation in the r value, the distribution of a variation in travel time depending on section lengths is shown in Table 2. Considering 90% of cumulative distribution in a variation in travel time, a variation in travel time on the Seoul-Suwon section, Seoul-Giheung section, and Seoul-Osan section is one minute, while a variation in travel time on the Seoul-Anseong section and the Seoul-Cheonan section is two minutes and about four minutes. A variation in travel time for a day not a mean value for a month may be higher. r is set as follows at m=0.17 and =45 by comparing the ratios of a variation in travel time depending on section lengths as the distribution of mean variation:

$$r = \frac{2}{1 + 2e^{-0.17(d-45)}} + 1$$

Section	Seoul to Suwon (17.2km)		Seoul to Osan (31.4km)		Seoul to Anseong (49.5km)		Seoul to Cheonan (69.4km)	
Variation (Minute)	Number of Vehicles	Distribution (%)	Number of Vehicles	Distribution (%)	Number of Vehicles	Distribution (%)	Number of Vehicles	Distribution (%)
0	1,708	63.5	1,194	44.4	754	28.1	487	18.1
1	950	35.3	1,249	46.5	1,139	42.4	848	31.6
2	26	1.0	218	8.1	517	19.2	570	21.2
3	3	0.1	20	0.7	175	6.5	294	10.9
4	1	0.0	4	0.1	58	2.2	183	6.8
5			1	0.0	22	0.8	121	4.5
6			2	0.1	6	0.2	63	2.4
7					4	0.1	32	1.2
8					6	0.2	30	1.1
9					4	0.1	20	0.8
10					1	0.0	9	0.3

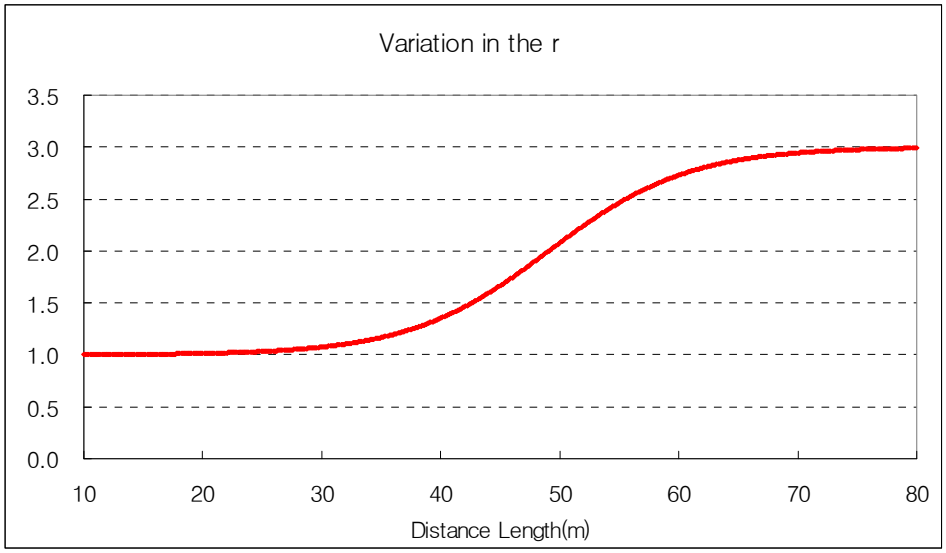
Total	2,688	100.0	2,688	100.0	2,688	100.0	2,684	100.0

<Table 2> Cumulative Distribution of a Variation in Travel Time Depending on Lengths

The r value of each section where the cumulative distribution is applied is shown in Table 3. As the section length gets shorter, the r and k values get lower and the level of smoothing gets higher.

Section	Suwon (17.2km)	Giheung (22.3km)	Osan (31.4km)	Anseong (49.5km)	Cheonan (69.4km)
r	1.0	1.0	1.1	2.0	3.0

<Table 3> Application of Value r



<Figure 6> Variation in Distance Coefficient (m=0.17, =45)

2) Setting of Allowable Amount of Variation

Since the effect of the introduction of Distance Coefficient r on section lengths is applied when Smoothing Constant k is calculated, an allowable amount of variation should be determined. Figure 7 denotes the exponential function of $y = e^x$. y-axis denotes Smoothing Constant k, and x-axis denotes the allowable amount of variation. Based on k=0.5 at x=ln(0.5)=-0.69, if the amount of variation, $\left|t_n - \hat{t}_{n-1}\right\}$ is higher than Allowable Amount of Variation q, the x value should be moved to the left to make the k value lower. Based on ln(0.5), one half of the travel time, $\left|t_n - \hat{t}_{n-1}\right| / q$, the row showing an effect on a variation in travel time and $1/r$, the row showing an effect on section lengths are introduced to calculate Smoothing Constant k as follows:

$$k = e^{\frac{In(0.5)}{r}\frac{\left|t_n - \hat{t}_{n-1}\right|}{q}}$$

In the formula above, ln(0.5) denotes the x value in the event of y=0.5 at $y = e^x$. In the event of $X = \frac{In(0.5)}{r}\frac{\left|t_n - \hat{t}_{n-1}\right|}{q}$ =-0.69=ln(0.5), k=0.5. This means that only one half of the amount of variation is accepted. Depending on the section length, a reference point of accepting only one half of the variation amount is applied differently.

When q is set to 10 minutes considering the variation distribution of travel time, r equals 1 on the Seoul-Suwon section. Depending on the variation amount of travel time, k is applied as follows:

i) When Amount of Variation $\left|t_n - \hat{t}_{n-1}\right|$ equals 10 minutes, $\frac{\left|t_n - \hat{t}_{n-1}\right|}{q}$ equals 1 and k equals 0.5.

The variation amount of travel time is applied by only half.

ii) When Amount of Variation $\left|t_n - \hat{t}_{n-1}\right|$ is higher than 10 minutes, $\frac{\left|t_n - \hat{t}_{n-1}\right|}{q}$ is higher than 1.

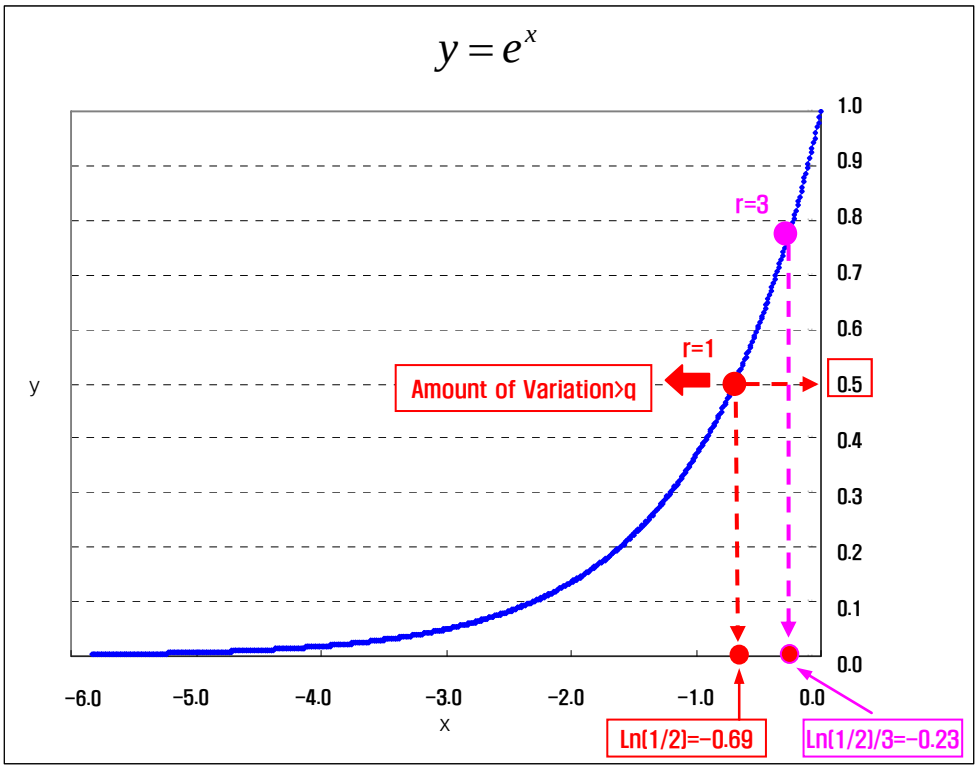
$In(0.5)\frac{\left|t_n - \hat{t}_{n-1}\right|}{q}$ is higher than ln(0.5).

The k value gets lower by moving to the left of the graph, and reflects a variation in travel time only a little.

iii) Since Amount of Variation $\left|t_n - \hat{t}_{n-1}\right|$ is less than 10 minutes and $\frac{\left|t_n - \hat{t}_{n-1}\right|}{q}$ is less than 1,

$In(0.5)\frac{\left|t_n - \hat{t}_{n-1}\right|}{q}$ gets lower than ln(0.5).

The k value gets higher by moving to the right of the graph, and reflects a variation in travel time much.



<Figure 7> Introduction of Allowable Variation Amount

Table 4 shows a variation in the k value depending on section lengths and the amount of variation at q=10.

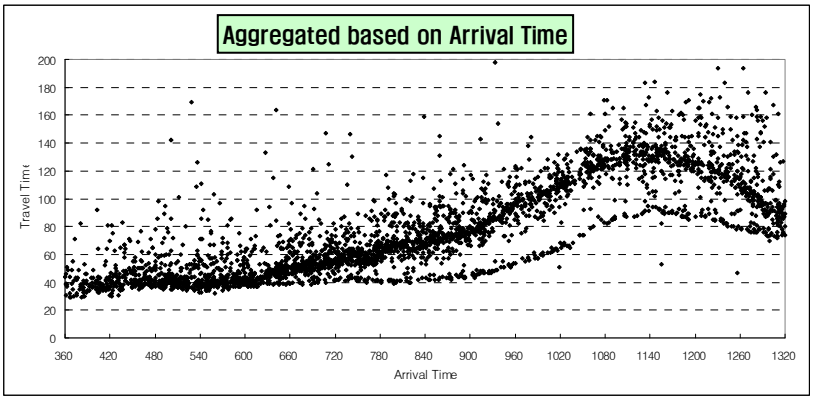
$\lvert t_n - \hat{t}_{n-1} \rvert$	$\frac{\lvert t_n - \hat{t}_{n-1} \rvert}{q}$	$\frac{In(0.5)}{r}\frac{\lvert t_n - \hat{t}_{n-1} \rvert}{q}$			k		
		r=1	r=2	r=3	r=1	r=2	r=3
5	0.5	−0.35	−0.17	−0.12	0.71	0.84	0.89
10	1.0	−0.69	−0.35	−0.23	0.50	0.71	0.79
15	1.5	−1.04	−0.52	−0.35	0.35	0.59	0.71
20	2.0	−1.39	−0.69	−0.46	0.25	0.50	0.63
25	2.5	−1.73	−0.87	−0.58	0.18	0.42	0.56
30	3.0	−2.08	−1.04	−0.69	0.12	0.35	0.50
35	3.5	−2.43	−1.21	−0.81	0.09	0.30	0.45

<Table 4> Value k Depending on the Section Length and the Amount of Variation

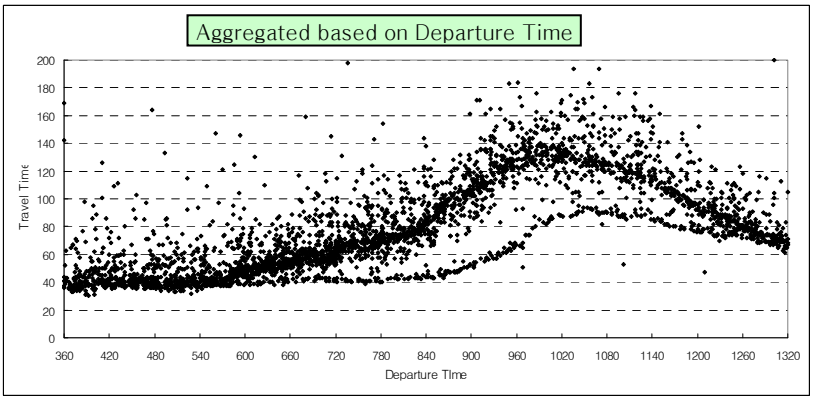
5. Application of the Model

When the pre-treatment model of the TCS data, which was developed in this paper, is applied by using the data on the Seoul-Cheonan section dated January 23, 2009, the results are shown

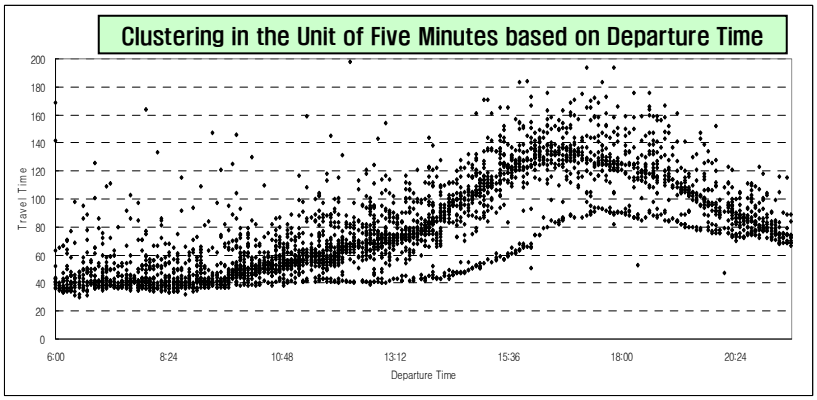
in Figures 8 to 11. The day was the Chinese New Year's Eve. Traffic to the direction of Busan on the Gyeongbu Expressway continued to increase by late afternoon. This confirms that the method proposed in this paper can reflect the decrease/increase pattern of travel time well.



<Figure 8> TCS Source Data



<Figure 9> Clustering in the Unit of Five Minutes Based on Start Time

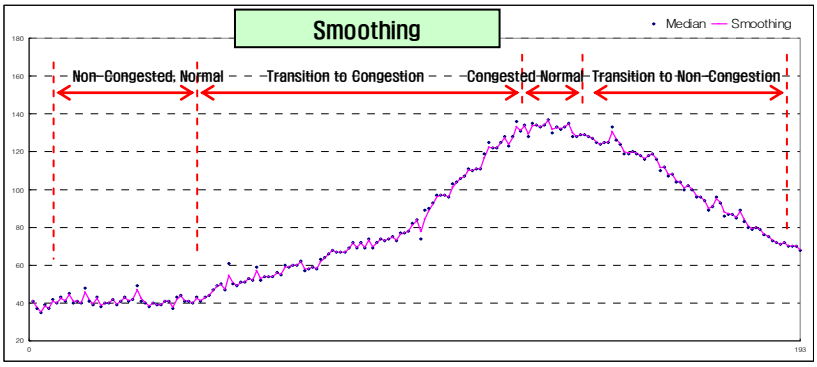


<Figure 10> Elimination of Outliers

When the TCS source data based on arrival time shown in Figure 8 are sorted out at five minutes of aggregate interval as shown in Figure 9, the increase pattern of travel time can be

found out to some degree; however, it can be verified that the distribution of travel times varies depending on start times. Thus, z_{cut} was set to different values depending on a variation of travel time, and the outlier was eliminated. The result is shown in Figure 3.10. The outlier was efficiently eliminated by setting z_{cut} to different values depending on a variation of travel time rather than 3.0. The representative value of travel time can be obtained as shown in Figure 3.10 by removing the extreme value through smoothing.

The traffic patterns on the Chinese New Year's Eve are a non-congested normal state, a transition to congestion, a congested normal state, and a transition to non-congestion as time goes by. Travel patterns can be properly assumed based on the TCS source data by using the pre-treatment method proposed in this paper. The representative value of travel time reflects travel patterns well at hours when a travel time sharply increases.



<Figure 11> Smoothing

6. Conclusion

In order for the TCS data to be of value as data on travel times, pre-treatment processes such as conversion the basis to a start time, the elimination of outliers, clustering, the selection of a representative value, and smoothing are required.

In this paper, the elimination variable of median absolute deviation was adjusted to decrease a deviation within the same aggregate interval and to minimize the effects of outliers when the representative value of the TCS is selected. As a result of the spatiotemporal, differential application of the elimination value by examining the characteristics of the coefficient of variation in travel times depending on section lengths and the circumstances of traffic delay and congestion, the gap between the upper and lower-limit confidence width of the representative value of travel time was able to be narrowed. In addition, a smoothing model for travel times was proposed by introducing a distance coefficient depending on section lengths and an allowable variation in travel times. As a result of applying the model, the estimated travel time turned out to be closer to the true value even at transition hours when a travel time sharply increased.

References

1.Kim Jae-jin, Roh Jeong-hyeon, and Park Dong-ju, Estimation of Link Travel Times Based on the Online Start Time Using VDSs, Vol. 24, the Journal of the Korean Society of Transportation, pp. 157-168, 2006.

2.Do Myeong-sik, Lee Hyang-mi, and Nam Gung-seong, Elimination of Outliers and the Development of an Algorithm for Incomplete Data using the TCS Data, Vol. 26, the Journal of the Korean Society of Transportation, pp. 241-250, 2008.

3.Lee Jeong-hui and Lee Yeong-in, A Paper of the Setting of Minimum Number of Samples to Present Travel Times on a Section, the 36th Autumn Seminar by the Korean Society of Transportation, pp. 458-462, 1999.

4.Jang Jin-hwan, Baek Nam-cheol, and Kim Seong-hyeon, Estimation of Dynamic Travel Times based on the AVI Data, Vol. 22, the Journal of the Korean Society of Transportation, pp. 169-175, 2004.

5.A Paper of the Improvement of Data on Travel Times on Expressways and of Using the Data, Korea Expressway Corporation, 2008.

6.Ashish, S., Piyushimita, T., Xioquon, Z., and Alan, F., Frequency of Probe Vehicle Reports and Variance of Arterial Link Travel Time Estimates, Journal of Transportation Engineering, ASCE, Vol. 123, No. 4, pp. 290-297, 1997.

7.Bajwa, S., Chung, E., and Kuwahara, M., Sensitivity Analysis of Short-term Travel Time Prediction Model's Parameter, 10th ITS World Congress, Madrid, Spain, 2003.

8.Hellinga, B. and Gudapati, R., Estimating Travel Times from Different Data Sources for Use in ATMS and ATIS, Proceedings of the ITE District 1 & 7 Joint Annual Conference held in Niagara Falls, Ontario, May 6, 2000.

9.Chi, X. and Reuy, L., Improving Arterial Link Travel Time Estimation by Data Fusion, 83rd TRB Annual Meeting, 2004.

10.Dion, F. and Rakha, H., Estimating Spatial Travel Time Using Automatic Vehicle Identification Data, 82nd TRB Annual Meeting, 2003.

11.Dion, F. and Rakha, H., Estimating Dynamic Roadway Travel Times Using Automatic Vehicle Identification Data for Low Sampling Rates, Transportation Research Part B, Vol. 40, No. 9, pp. 745-766, 2006.

12.Emam, E. and Al-Deek, H., Utilizing a Real Life Dual Loop Detector Data to Develop a New Methodology for Estimating Freeway Travel Time Reliability, 85th TRB Annual Meeting, 2005.